F79MA 2025-26
Assessed Project 2

Project description

Biologists are often interested in the number of counts of a genetic mutation in a fixed length of an RNA sequence.

It has been observed that the number of counts, X, can often be effectively modelled by a negative binomial distribution with probability mass function

$$\Pr(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \qquad x = k, k+1, \ldots$$

where $0 < p < 1, k = 1, 2, 3, 4, \ldots$ are suitably chosen parameters. Note that k and p may not have natural interpretations.

The aim of this project is to perform a Bayesian analysis of some given count data for genetic mutations.

Instructions on code and report

In this project, you are provided an R script, `F79MA_Assignment2.R` (available on Canvas) to get you started with computational aspects of this analysis. In the R script, you should set the variable `last_four_digits` to the last 4 digits of your Heriot-Watt student ID number. For example, if your student ID is H12345678, then you should set `last_four_digits = 5678`. **Failure to do so will result in 0 marks being awarded.**

You have also been provided with a dataset, `count_data.csv`, containing some count data. Please ensure you set your working directory in R to be wherever you have saved the file `count_data.csv`. Provided you have done this, the R script will save the full data to a variable called `Full_Data` and also select a sample from the count data called `My_Data` which you will work with.

Aside from the variable, `last_four_digits`, do not change any of the other code above the line reading 'Please insert your R code after this line.' **Failure to do so may result in 0 marks being awarded.**

You must submit, using the links on Canvas: (a) your written report, as a pdf file or as a Word document; and (b) an R script containing code that calculates the required quantities and produces any plots you include. Before submitting, you must complete the Standard Declaration of Student Authorship quiz on Canvas.

Tasks

1. It is common to assume that count data in many contexts outside of biology can be modelled via a Poisson distribution.

   Explain in the language of Bayesian statistics why modelling count data in terms of a Poisson distribution with rate λ which is itself a Gamma distributed random variable could lead to the count data being modelled via a negative binomial distribution.

   Provide a reason why this may be more appropriate than directly using a Poisson distribution.

   Note that you can make use of results from lectures and tutorials. [2 marks]

It is believed that the count data `My_Data` you have been provided with can be modelled as independent and identically distributed realisations $x_1, \ldots, x_{1000}$ from a negative binomial distribution with parameter $k = 5$ and unknown parameter p.

2. The maximum likelihood estimate $\hat{p}$ for the parameter p in such a scenario is given by

$$\hat{p} = \frac{k}{\bar{x}}$$

   where $\bar{x}$ is the sample mean. Calculate the maximum likelihood estimate $\hat{p}$ for `My_Data` and store the value in your code as `Quantity1`. [1 marks]

You will now perform a Bayesian analysis on the provided count data.

3. Derive the Jeffreys' prior for the parameter p for a sample of size $n = 100$ from a negative binomial distribution with other parameter $k = 5$.

   Comment on the resulting prior distribution. [2.5 marks]

4. Derive the posterior density for the parameter p using the prior from Task 3 and the data from `My_Data`.

   Calculate the posterior mean and store it as `Quantity2` in your code. Comment on your results. [3 marks]

5. Plot the densities of the prior and posterior distributions, and comment on their shapes. [2 marks]

6. Using your analysis, derive the posterior predictive distribution $\pi(z|\underline{x})$ for an unseen count z.

   Calculate the value of $\pi(Z = 5|\underline{x})$ and store it as `Quantity3` in your code. [2.5 marks]

7. Simulate 10000 realisations from the predictive distribution from Task 6 and plot a histogram of your results.

   Plot a histogram of the data in `Full_Data`.

   Discuss the performance of the predictive distribution. [3 marks]

The Report

Your findings should be presented in the form of a report for your line manager, who is also part of the statistical modelling team but has not worked directly on this problem.

Your report should:

- have a clear and logical structure; include an introduction and clearly stated conclusions that can be understood by any numerate scientist;
- include detail of your mathematical calculations so that your results could be reproduced by another statistician;
- include clearly labelled and correctly referenced tables and diagrams, as appropriate;
- include citation and referencing for any material (books, papers, websites etc) used. When possible, use reliable sources, produced by respected and well-known authors, published by recognised publishers and associated with well established government, academic, or educational institutions. Note that some webpages, YouTube videos, blog posts, and Wikipedia pages might include errors.

Your report should be produced using a word-processing program, such as Microsoft Word, or with LaTeX, and uploaded as a single file in PDF format. All equations should be produced digitally (typeset) by using an equation editor (we recommend to avoid scanned or handwritten equations). You may use the Microsoft Word template provided, but this is not compulsory.

The report has a **maximum page limit of eight (8) pages** (equivalent to 4 double-sided A4 sheets, 11-point font).

A total of **4 Marks** is available for these aspects of your report. This will be marked according to the rubric given in the Appendix.

[Total: 20 Marks]

Notes

- **Errors in code:** Three marks are for correctly carrying out calculations in R (i.e. calculating the correct values for the Quantities), and this will be assessed by running your R code and comparing the results with the verified correct output. You must therefore ensure that you have defined variables as above (i.e. **exactly the same naming convention with capital Q and no underscore between Quantity and the number**) and ensure your code runs without generating errors. **Failure to do so may result in 0 marks being awarded**. You can check your code runs correctly by running it through the file `TestSubmissionAssignment2.R`.

- **Use of libraries and good practice with code:** Please do not give references to directories on your own computer (e.g. specifying a directory with setwd) or make use of any R library aside from the 'ggplot2' library (you do not need to use this library). Your code should be commented appropriately to ease with checking.

- **Questions for the Lecturers**: If you have any questions on the assignment, please ask in the discussion board rather than e-mailing lecturers directly. That way all students will see the same responses. The discussion board will be closed at 9:00 am UK time = 1:00 pm UAE time = 5:00 pm Malaysia time on Monday 24th November 2025.

- **Collaboration and Plagiarism**: Students are allowed to discuss the methods used with other students, but your submitted project must be **all your own work**. Particularly please note the following:

  - Coursework reports must be written in a student's own words and any code in their coursework must be their own code. If some text or code in the coursework has been taken from other sources, these sources must be properly referenced.
  - Failure to reference work that has been obtained from other sources or to copy the words and/or code of another student is plagiarism and if detected, this will be reported to the School's Discipline Committee. If a student is found guilty of plagiarism, the penalty could involve voiding the course or worse.
  - Students must never give hard or soft copies of their coursework reports or code to another student. Students must always refuse any request from another student for a copy of their report and/or code.
  - Sharing a coursework report and/or code with another student is collusion, and if detected, this will be reported to the School's Discipline Committee. If found guilty of collusion, the penalty could involve voiding the course.

More information on plagiarism can be found here

- **Submission**: The report and the code should be submitted through Canvas using the appropriate links. It should not be submitted by email or handed in as a paper copy. You should also complete the Student Declaration of Authorship quiz.
- **Deadline**: The deadline for submission is **9:00 am UK time = 1:00 pm UAE time = 5:00 pm Malaysia time** on **Tuesday 25th November 2025** (Week 12); projects may be submitted early. Late projects will be marked in line with the University Coursework Policy, i.e. a standard deduction of 30% if submitted within 5 working days and 0 marks awarded if submitted later. You can, however, submit a mitigating circumstances applications, but these may only be applied at the end of the semester. No individual extensions are allowed.
- **Feedback**: Feedback on this assignment will only be available after the exam period ends.
- **Assessment**: This coursework will contribute 20% to the final mark for the course.

Appendix: Rubric for marking of the report

The five marks available for the exposition of your report will be awarded according to the scale below:

Between 0 and 1 marks will be awarded for	• Lack of clear and logical structure • Conclusions missing or not suitable for a non-statistician • Statistical calculations and methodology not clearly set out for the reader • Tables and figures unclear, badly labelled or not correctly referred to • Sources used not clearly referenced
Between 2 and 3 marks will be awarded for	• Clear and logical structure including Introduction and Conclusion • Conclusions generally suitable for a non-statistician • Statistical calculations and methodology generally set out clearly for the reader • Tables and figures often clear and correctly referred to • Sources used clearly referenced
4 marks will be awarded for	• Clear and logical structure including Introduction and Conclusion • Conclusions suitable for a non-statistician • Statistical calculations and methodology set out clearly for the reader • Tables and figures clear, correctly referred to and easy to interpret • Sources used clearly and correctly referenced